DOCUMENT RESUME

ED 376 214 TM 022 351

AUTHOR Baxter, Gail P.; And Others
TITLE Cognitive Analysis of a Science Performance Assessment. Project 2.1 Designs for Assessing Individual and Group Problem Solving. Assessing the Validity of Existing Assessments of Problem-Solving Performance in Science: A Taxonomy of Cognitive Processes.
INSTITUTION National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.; Pittsburgh Univ., Pa. Learning Research and Development Center.
SPONS AGENCY National Science Foundation, Washington, D.C.; Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Jun 94
CONTRACT ESI-90-55443; R117G10027
NOTE 28p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Cognitive Processes; Cognitive Psychology; *Educational Assessment; Educational Practices; *Elementary School Students; Evaluation Methods; Grade 4; Grade 5; Intermediate Grades; Interviews; Pilot Projects; *Student Evaluation; Test Reliability; Test Use; Test Validity; *Thinking Skills
IDENTIFIERS *Performance Based Evaluation

ABSTRACT

The degree to which performance assessments meet their dual mandate to evaluate student learning and inform instructional practice is not adequately addressed through traditional concerns for reliability and validity. A possible approach is suggested for examining the cognitive activity students engage in during a performance assessment. The approach is demonstrated with the "Mystery Powders" classroom-based assessment being piloted by several large school districts. Thirty-seven fourth- and fifth- grade students were interviewed while they conducted an investigation to determine properties of various powders. Interview protocols and observations were analyzed, and high and low scorers were described. Results indicate that although performance scores and general understanding were generally low, high scorers could be distinguished on several characteristics. Results support the viability of the approach for analyzing the extent to which performance assessments measure higher-order thinking. Implications for instructional practice are considered. Seven figures and four tables are included. (Contains 7 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED376214

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

National Center for Research on
Evaluation, Standards, and Student Testing

Final Deliverable – June 1994

Project 2.1 Designs for Assessing Individual
and Group Problem Solving

Assessing the Validity of Existing Assessments of Problem-Solving Performance in Science: A Taxonomy of Cognitive Processes

Cognitive Analysis of a Science Performance Assessment

Robert Glaser, Project Director
CRESST/LRDC, University of Pittsburgh

U.S. Department of Education
Office of Educational Research and Improvement
Grant No. R117G10027 CFDA Catalog No. 84.117G

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

TM022351

BEST COPY AVAILABLE

The work reported herein was supported under the Educational Research and Development Center Program, cooperative agreement number R117G10027 and CFDA catalog number 84.117G, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

COGNITIVE ANALYSIS OF A SCIENCE PERFORMANCE ASSESSMENT[1]

Gail P. Baxter and Anastasia D. Elder
CRESST/University of Michigan

Robert Glaser
CRESST/LRDC, University of Pittsburgh

Abstract

The degree to which performance assessments meet their dual mandate—evaluating student learning and informing instructional practice—is not adequately addressed in traditional concerns for reliability and validity. New forms of assessments demand new forms of evaluation, evaluation that documents the nature and extent of student thinking and reasoning required for optimal performance. In this paper we suggest one possible approach for examining the cognitive activity students engage in during a performance assessment. We demonstrate the utility of this approach with "Mystery Powders," a classroom-based assessment currently being piloted by several large school districts. Thirty-seven fourth- and fifth-grade students were interviewed while they conducted an investigation to determine the properties of various powders. Interview protocols and observations were analyzed with respect to several characteristics of proficient performance such as planning, monitoring, solution strategy, and explanations. Students with differing levels of competence (e.g., high and low scorers) were

[1] This assessment task was developed through a grant from the National Science Foundation (ESI 90-55443) to the first author and her colleague, Richard J. Shavelson, University of California, Santa Barbara. The work reported herein was supported under the Educational Research and Development Center Program, cooperative agreement number R117G10027 and CFDA catalog number 84.117G, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education. Special thanks to Stephen Druker, University of California, Santa Barbara, for his considerable help in the design of the Mystery Powders assessment and scoring system. Thanks also to Jasna Jovanovich, University of Illinois at Urbana-Champagne, Karen Malhiot, Chicago Public Schools, and Tim Breen, University of Michigan, for their help in data collection. Last, but not least, we wish to gratefully acknowledge the support of the teachers, students, parents and administrators of two school districts without whom this study would not have been possible.

described in terms of these characteristics. Results indicate that although performance scores and general understanding were generally low, high scorers could be distinguished from low scorers on several characteristics. High scorers provided a more complete plan for approaching the task, were more strategic in their problem-solving approach, engaged more frequently in self-monitoring activity, and generated better explanations of content-related concepts than low scorers. The results suggest the viability of this approach for analyzing the extent to which performance assessments measure higher order thinking. Moreover, characterizing student performance with respect to dimensions of competent performance has direct implications for instructional practice.

Alternative forms of assessment have been proposed as a major impetus for educational change. These assessments are intended to evaluate student understanding and provide models of performance that educational practice should foster in all students. Despite this auspicious mandate for assessment programs, criteria for evaluating their effectiveness continue to revolve around traditional concerns for reliability and validity. "A better research base is needed to evaluate the degree to which newly developed assessments fulfill expectations. Claims that performance assessments measure higher order thinking skills and deep understanding, for example, require detailed cognitive analysis" (Baker, O'Neil, & Linn, 1993, p. 1216). This detailed cognitive analysis should examine the underlying processes that define the nature of the problem-solving activities that contribute to proficient performance, and illustrate the kind of performance actually elicited from students in alternative assessment situations. A strong positive relationship between performance score and processes of thinking and reasoning provides evidence for the claims that these assessments measure higher order thinking (Baxter, Glaser, & Raghavan, 1993). However, procedures for carrying out this type of analysis are not well established. In this paper we offer one possible approach for consideration, an approach that grows out of research on the nature of competent performance in a subject matter.

Characteristics of student performance that develop as students display increasing proficiency in problem solving and higher order thinking have been suggested by the literature on expertise (see Chi, Glaser, & Farr, 1988). Proficient or competent students are characterized by: (a) integrated and well-

connected knowledge that fosters their ability to reason and make inferences with what they know, (b) usable knowledge that extends beyond mere factual information and allows information to be applied in appropriate situations, (c) a set of proceduralized or automatic skills which free up time for higher order thinking, (d) the ability to effectively represent the meaning of a problem and plan an approach before employing a solution strategy, (e) the ability to explain the reasons for particular actions, and (f) a repertoire of well-developed self-regulatory skills used to monitor and control performance. Because these characteristics indicate proficiency with a subject matter, they provide a useful framework for analyzing students' thinking and reasoning in an assessment situation. These characteristics should not be viewed as a rigid list of criteria to which all assessments must conform. Rather they serve as a general framework for guiding cognitive analyses of a performance assessment. In some assessment situations, particular characteristics may be more salient than others or more difficult to infer.

This paper documents the kind and level of cognitive activity students engage in while conducting a performance assessment designed to provide feedback to students and teachers about students' level of understanding after completing an elementary science unit. The unit, Mystery Powders, was first developed in the 1960s as part of the Elementary Science Study (ESS). Although individual programs have adapted the ESS unit for their own use, the intent of the unit and the sequence of activities have been maintained. Because this unit is taught as part of many inquiry-based elementary science programs, an assessment was developed that teachers could administer and score in their classrooms. Currently this performance assessment is being piloted in several urban and large suburban school districts.

To evaluate the assessment, detailed verbal protocols of students' performances were collected, and analyses were guided by the characteristics of proficient performance described above. Analysis of these protocols in conjunction with observations of students' performances provide a basis for linking performance scores with level and kind of reasoning and understanding (e.g., Baxter et al., 1993). It is expected that students who perform well on the assessment will display some characteristics of proficient performance (ability to plan, explain, draw inferences, systematically solve problems, and monitor their own performance). Moreover, it is expected that

the quality of these characteristics will vary with performance score. Meeting these expectations provides evidence that this assessment necessitates higher order thinking skills and the engagement of the appropriate cognitive activity for adequate performance.

Hands-On Science Teaching and Learning

As science education moves to embrace the constructivist notions of learning, hands-on science programs are seeing renewed popularity in elementary classrooms. Typically, in these inquiry-oriented programs, students study three to five science units during a school year in a sequence prescribed by the teacher and/or the school district. Each kit-based unit is designed to promote an understanding of a few concepts and processes through the completion of an eight- or nine-week sequence of activities. Some activities are of an exploratory nature, some involve planned investigations, and some involve application to "real world" situations. In general, students work with a partner or in small groups to generate questions and hypotheses, plan strategies, conduct activities and draw conclusions. Teachers can facilitate development of student understanding by providing meaning to the activities, setting the occasion for informative or evaluative comparisons, and drawing attention to the principles underlying a set of procedures (Schauble, Glaser, Duschl, Schulze, & John, in press).

For example, the Mystery Powders unit engages students in systematic investigation of the properties of substances. Students study five simple, white powders—salt, sugar, baking soda, cornstarch, and plaster of paris. The purpose of the unit is to help students understand: (a) that each powder has a unique set of properties, some physical and some chemical; (b) that particular properties are more salient than others and therefore more reliable for identification purposes (e.g., iodine turns black when mixed with cornstarch but not with other powders); and (c) that a combination of confirming and/or disconfirming evidence may be required for the identification of a powder(s).

To develop this understanding, teachers guide students through systematic investigation of each of the powders. Students document in their science journals the tests (adding water, vinegar, iodine, or heat), observations, and information from sensory input (smell, touch, taste, and observation with a hand lens) for each of the five powders. Comparing and

contrasting the accumulated information draws attention to the differential reliability of each method for distinguishing one powder from another. For example, iodine turns purple/black with cornstarch but yellow or orangish with all the other powders, and vinegar fizzes with baking soda but not with any of the other powders. Sugar melts and becomes caramel-like when heated. Salt has a unique crystal structure (uniform, cube-shaped) when viewed under a hand lens. Plaster of paris becomes hard when mixed with water and, unlike the other powders, will remain hard when water is mixed with it again.

Mystery Powders Assessment

The development of the Mystery Powders assessment was guided by a team of researchers, scientists, and teachers. The intent was to develop a performance test that teachers could administer and score to evaluate students' understanding and to inform instructional practice. For the Mystery Powders assessment, students are asked to identify the contents of six bags given a list of five possible options; some bags contain individual powders, some contain two powders. Students engage in an iterative sequence of hypothesis generating, testing out hypotheses, evaluating observations, and drawing conclusions until a solution is reached. Students use their journals from science class as a resource when completing the assessment.

Two scores are assigned to each student, one for the answer (identification of the powder[s]) and one for the evidence. The identification score provides information solely about accuracy of the answer—a more traditional way of assessing students. The evidence score provides information about the tests students attempted and the observations they deemed important as well as their understanding of necessary and sufficient evidence to support their conclusions. For example, if a student's identification score is low yet the evidence score is high, then the student may be having problems with drawing conclusions based on his or her evidence. Using two distinct scores provides information on the nature of students' difficulties.

Teachers administer this assessment to all students and score the performance in an effort to gauge the extent to which students understand the key concepts and processes of the Mystery Powders unit. Further, the performance scores provide systematic feedback to teachers about the

strengths and needs of their students. In this paper we document the cognitive activity students engaged in during the conduct of this assessment and examine the relationship between performance score and the nature and extent of cognitive activity.

Method

Subjects

The Mystery Powders assessment was administered to 37 fourth- and fifth-grade students within one week of completing an eight-week, activity-based unit of study on the properties of substances. Students represented a diverse range of ethnic/cultural backgrounds, and males (n = 20) and females (n = 17) were equally represented. All students lived in an urban or large suburban school district where they participated in districtwide, inquiry-based science programs for two or more years. Students were chosen by their respective teachers in each of six different classrooms with the express purpose of ensuring a range of science ability. Interviewers were unaware of the teachers' rankings of students' science ability.

Instrumentation

Task. Students were asked to identify common white powders such as salt, baking soda, and cornstarch contained in each of six bags—some individually (e.g., baking soda), some in combination (e.g., baking soda and cornstarch). The possible contents of the bags were clearly conveyed to the students—baking soda, cornstarch, baking soda and salt, cornstarch and salt, and baking soda and cornstarch (see Figure 1). Each of the six bags contained one of the possible options; two of the bags contained the same thing. Students were provided with iodine, water, vinegar, and a hand lens to conduct their investigations. Students could also consult their science journals where they recorded the results of their investigations of each of the powders during science class. After conducting tests of the six bags of powders, making observations, and recording their notes, students were asked to summarize their findings in a table of results and conclusions (see Figure 2).

Find out what is in each of the bags 1, 2, 3, 4, 5, and 6. Use any of the equipment on the table to help you determine what is in each bag. **Each bag has <u>one</u> of the "Mystery Powders" listed below.**

Baking Soda	*(D)*
Cornstarch	*(A)*
Cornstarch and Baking Soda	*(A and D)*
Baking Soda and Salt	*(D and C)*
Cornstarch and Salt	*(A and C)*

NOTE: Two of the bags will have the same thing. All of the others will have something different.

==========

<u>Keep notes</u> on what test(s) you did and what you observed as you conduct your investigation. You have room on the following pages. Use your notebook from science class to help you determine what each powder is. When you think you know what is in a bag, record your results and conclusions in the table on the last page.

Figure 1. Mystery Powders assessment.

It should be noted that students may have referred to the powders by the letters A, B, C, D, E, and F because during the course of instruction the intent was not to identify the powders by name but to focus on observing those properties that distinguish one powder from another. As such, some students do not know the "names" of the powders and refer to them only by a letter.

RESULTS AND CONCLUSIONS

Look at the tests and observations you made today.
Check your observations with the notebook you kept during science class.
Fill in the table below.

Mystery Powder	What's inside the bag?	What test(s) told you?	How did you know? What happened?
1	*Cornstarch & Baking Soda*	*Vinegar and iodine.*	*When I added vinegar it fizzed, and when I added iodine it hardend and turned black.*
2	*Baking Soda & Salt*	*Vinegar and iodine.*	*When I added vinegar it fizzed, and when I added iodine it became two layers.*
3	*~~Cornstarch~~ & Baking Soda salt.*	*Vinegar and iodine and water.*	*When I added vinegar it turned cloudy, When I added iodine it also turned cloudy and milky.*
4	*Cornstarch*	*Water, Vinegar, iodine.*	*When I added water it harden, when I added vinegar it really hardened, and when I added iodine it turned black.*
5	*~~Cornstarch~~ & salt Baking Soda*	*Vinegar and iodine*	*When I added vinegar it started to turn cloudy, when I added iodine it turned black.*
6	*Cornstarch*	*Vinegar, iodine, water.*	*When I added Vinegar it fizzed. When I added iodine it turned black. When I added water, it made it easier to compare.*

Figure 2. Results and conclusion page of Mystery Powders assessment.

Scoring. Students' results and conclusions were read, and the responses were transcribed on the score form. Credit was given for correctly identifying the substance in each bag and for the tests/observations that led to that conclusion. For the identification score, students received 1 point for correctly identifying the contents of a bag, zero points for incorrect identification. Students did not receive partial credit for identifying one powder out of two powders in a bag. For example, if students indicated that there was cornstarch in bag 1, they received 0 points since bag 1 contains cornstarch and baking soda. The points were summed over all six bags for a total answers correct score of 0 to 6 (see left-hand column of Figure 3).

For the quality of evidence score, students received 1 to 4 points for providing appropriate support for their answer. Tests and observations constitute evidence, and students were evaluated on the quality of the evidence they provided (see bottom of Figure 3). Quality was dependent to some extent on the contents of the bag. When the "Mystery Bag" contained two powders (e.g., cornstarch and baking soda), students who confirmed or ruled in the presence of each powder (provided complete evidence for both powders) received 4 points. Note that the confirming tests are those indicated with a black box on the score form. Students who provided complete evidence for one powder but incomplete evidence for the other powder received 3 points. Observations that provided partial or incomplete evidence are indicated with a line drawn under them. Students who provided complete evidence for one powder but no evidence for the other powder received 2 points. Students who provided incomplete evidence for both powders or inadequate evidence received 1 and zero points, respectively (see Figure 3). For example, if students reported tests (e.g., vinegar, iodine) without corresponding observations (fizzed, turned black), zero points were awarded.

When only one powder was present (e.g., baking soda), students must confirm or rule in the presence of that one powder and disconfirm or rule out the presence of all other powders which may be in combination with that powder (cornstarch, salt) to receive 4 points. Note that the disconfirming tests are indicated by white boxes on the score form. As was the case with two powders, points are awarded based on the quality of evidence. The less complete the evidence confirming the presence of one powder and

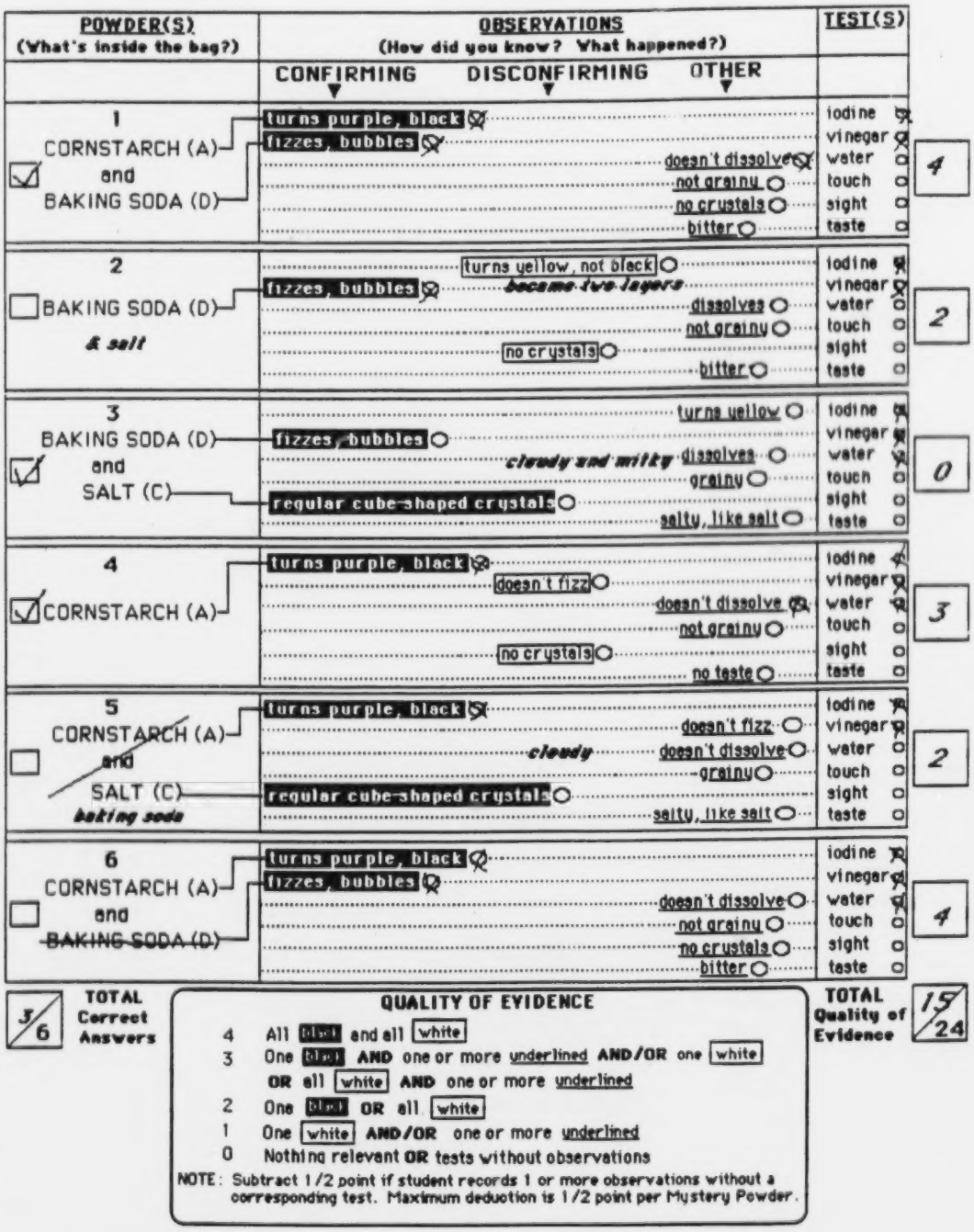

Figure 3. Mystery Powders score form.

BEST COPY AVAILABLE

disconfirming or ruling out the presence of a second powder, the lower the score (see Figure 3).

If a student provided observations without tests, one-half point was subtracted from his or her evidence score. For example, consider a student who reports testing with vinegar and reports two observations (fizzed and turned black) for bag 1 (baking soda and cornstarch). Fizzed with vinegar confirms the presence of baking soda and turned black with iodine confirms the presence of cornstarch. This student would initially receive 4 points. However, the student failed to report the test which led to the observation "turned black." One-half point is subtracted for a final score of 3 1/2 points. In scoring then, reporting observations is more important than reporting tests.

Consider the student responses in Figure 2. For bag 1, the student receives 1 point for the answer (identified the powders as cornstarch and baking soda). As evidence, the student reported fizzing with vinegar and turned black with iodine. These tests and observations are considered complete evidence for both powders. Therefore the student receives 4 points for evidence (see Figure 3).

As a second example, consider the student's response for bag 2 (see Figure 2). The student incorrectly identifies the contents of the bag as baking soda and salt (identification score = 0). As evidence, the student noted fizzing with vinegar and "became two layers" with iodine. The student did not provide evidence to rule out cornstarch (iodine turns yellow, not black) or salt (no crystals). These tests and observations are required for complete evidence (4 points). The student receives 2 points for providing complete evidence for one powder (baking soda; see Figure 3).

Procedure

Students were interviewed and audiotaped individually, while they conducted the assessment task, that is, while they tried to identify the powder(s) in each of the six bags. Directions were read aloud to all students and all equipment was introduced (vinegar, iodine, water, hand lens, spoons, stir sticks, cups). After hearing the instructions, but before the students began, they were asked if they understood the task and to explain what they were being asked to do. Next, they were asked about their plans for completing the assessment ("Can you tell me how you're going to go about it?"). While

conducting the assessment, students were prompted with questions to simulate a think-aloud procedure (see Ericsson & Simon, 1984). Students were encouraged to verbalize their strategies and articulate their reasoning while carrying out the assessment (e.g., "Why are you adding iodine?") and when drawing conclusions (e.g., "How do you know it's baking soda and salt?"). Space was provided for students to keep notes and record their observations as they conducted the tests. Interviewers recorded the procedures and strategies students used. For example, interviewers noted the sequence of tests conducted on each bag. They also noted when students referred to their science journals or the list of possible contents for each of the six bags. After completing all tests and observations students were prompted to look at their notes, consult their science journals, and summarize their findings in a table of results and conclusions. Interviewers then asked a final question aimed at their general understanding of the unit: "Can you determine the contents of each bag by using water only?"

Results and Discussion

Transcriptions of audiotaped interviews, interviewers' written observations of students' strategies and activities (e.g., referred to science journal to check observations, hypotheses, and conclusions), and students' performance (evidence) scores served as data. Analyses were guided by dimensions of proficient performance. Specifically, we examined the relationship between students' performance scores and their ability to: (a) explain principles underlying task performance, (b) generate a knowledge-based plan for approaching the task, (c) utilize a principled problem-solving approach or strategy, and (d) monitor their performance. It was expected that students who scored high on this assessment would plan, explain, systematically solve the problem, and monitor at a level that was qualitatively different from that of students who scored low. In the following sections we describe: (a) the nature of student thinking and reasoning with respect to each of the aforementioned characteristics, and (b) the correspondence between students' scores and the cognitive characteristics they display. Distinctions between high- and low-scoring students on each of these characteristics provide evidence to support inferences that this assessment taps relevant higher order thinking.

Performance Scores and General Understanding

On average, students correctly identified 2.6 (*sd* = 2.09) of the 6 bags. Mean evidence score on the task was 9.11 (*sd* = 4.96) out of a possible 24 points. Although possible evidence scores could range from 0 to 24 (6 bags x 4 points per bag), the maximum score obtained in this sample of students was 18 (see Table 1). Indeed, the scores were quite low, with more than two-thirds of the students scoring 11 or less.

Recall the purpose of the unit was to develop students' understanding of the properties of substances and the types of tests that could be used to indicate those properties and thereby identify the substance. As a measure of the extent to which students developed this general understanding we asked: "Could you identify all the powders by testing with water only?" If students understand that various tests are differentially effective for the identification of each of the powders, then they would recognize that one test would not adequately distinguish among the powders.

Students' responses were evaluated on the quality of their explanation and categorized according to one of three levels: inadequate, partial, good (see Figure 4). Each level is distinguished by the completeness and coherence of

Table 1

Distribution of Evidence Scores

Range of evidence scores	Number of students	Percentage of students
20-24	0	0
16-19	5	14
12-15	7	19
8-11	11	30
4-7	9	24
0-3	5	14
Total	37	101[a]

[a] Percentages do not add to 100 due to rounding.

Inadequate:	Incorrect response or "I don't know." *"Yes it could get watery and you'd say cornstarch."*
Partial:	Correct response with specific example. *"No, because when I put vinegar on powders it did make bubbles, but when I put water on it, it didn't make that many bubbles, it just sank down. So that's why I think no."*
Good:	Correct response with general explanation/description. *"No, because they're different powders. They don't do the same things. And there are other things that would tell me what they are and water wouldn't.*

Figure 4. Quality of response to content knowledge.

students' explanations: (a) inadequate responses include restating the question, "I don't know" statements or incorrect responses; (b) partial responses describe a particular occurrence or specific example; (c) good responses provide generalized explanations.

Consistent with the low performance scores reported above, two-thirds of the students provided a partial or inadequate explanation (see Table 2). Students who scored high provided good explanations that reflected their understanding of the differential effectiveness of each of the various tests. In contrast, all students who scored low (0-3), except one, provided inadequate explanations. Students with scores ranging from 4 to 7 provided partial explanations at best.

Plan

Proficient students provide a plan that guides their solution strategy, a plan based on their representation of the task and the principles on which performance is dependent. Less proficient students, in contrast, do not generate an adequate representation of the task because of a lack of general

Table 2

Proportion of Students Generating Explanations of Various Quality

Range of evidence scores	Number of students	Quality of Explanation		
		Inadequate	Partial	Good
20-24	0	—	—	—
16-19	3	—	—	1.0
12-15	7	.1	.5	.5
8-11	10	.3	.5	.2
4-7	6	.5	.5	—
0-3	3	.7	—	.3
Total	29[a]			

[a] Data are not available for 8 of the 37 students

understanding of the relationship between the powders, the tests and the observations, and how these might be used to identify the contents of each bag. As a consequence these less proficient students begin solving the task without thinking through the entire solution. It was expected, then, that those students with higher scores would provide more complete plans; those with low scores would provide less complete plans or no plan at all.

Before beginning the investigation, but after students described the task, they were asked "Can you tell me how you are going to go about it (your investigation)?" Plans were categorized with respect to one of four levels based on the completeness with which the task was represented and addressed (see Figure 5).

Given the generally low scores on this assessment, it is not surprising that none of the students gave an elaborated plan that would reflect an understanding of the properties of the powders and what would constitute necessary and sufficient evidence for the identification of each powder. For example, students did not say they would test with vinegar to indicate which bags had baking soda, with iodine to indicate which bags had cornstarch, and with a hand lens to indicate which bags had salt.

Level 1: Restates problem or procedure.
"test them" or *"check them."*

Level 2: Names tests.
"By observing it and putting water or vinegar or iodine in it. And that's how I'll find out."

Level 3: Names tests and reactions specific to one or more powders.
"Okay I'm going to open it and first try the vinegar on it... I'll know it's baking soda because of the bubbles, so and like cornstarch feels soft."

Level 4: Names tests and need to compare to prior results. May provide specific example of test and reaction.
"With my book...I took notes when we used to do our test. I'm gonna find out with the vinegar, water, and iodine.... Like when the baking soda fizzed, with the vinegar that's what it told me because in my book I wrote that when you put vinegar in the baking soda it fizzes, and with the taste it tastes funny, like salty."

Figure 5. Quality of plans.

Approximately three-fourths of the 37 students generated a Level 1 or 2 plan characterized by their focus on the procedures or materials (test, use vinegar) without reference to what the tests might tell them or how the information might be used to identify the powders. Four of the students displayed some general understanding that powders have properties and that each of the tests is a differentially effective method for observing those properties. Their plans (Level 3) explicitly mentioned the relationship between a test/observation and a particular powder. The four students (11%) who provided a Level 4 explanation stated how the investigations they had done in class would serve as the basis for comparing their observations and checking their conclusions.

Proportions of students providing each level of plan (1-4) varied with evidence score (see Table 3). Those with the lowest evidence scores (0-3) were more likely to provide a Level 1 plan, and those with scores 4-11 were most likely to provide a Level 2 plan. Students with the highest scores (12-19) varied in the quality of the plan they provided.

Strategy

The Mystery Powders unit is intended to develop students' understanding of the distinguishing properties of powders, which on the surface look quite similar (i.e., white), and the utility of using these properties for identification purposes. This understanding will be reflected in the strategy or approach students use to determine the contents of each bag. For example, students may examine the list of contents (baking soda, baking soda and cornstarch, baking soda and salt, cornstarch, cornstarch and salt) and recognize that certain tests would be more informative than others. Vinegar will indicate if baking soda is in the bag, iodine will indicate if cornstarch is in the bag, and examination of the powder under a hand lens will indicate if salt is present because of the regular cube-shaped appearance of salt crystals. Testing with vinegar, and iodine, and observation with a hand lens would be the most efficient strategy to use to identify the contents of the six bags. Relying on less

Table 3

Proportion of Students Within Given Score Range Generating Various Quality Plans

Range of evidence scores	Number of students	Quality of Plan			
		1 Low	2	3	4 High
20-24	0	—	—	—	—
16-19	5	—	.4	.4	.2
12-15	7	.3	.3	.1	.3
8-11	11	.1	.8	.1	—
4-7	9	.4	.6	—	—
0-3	5	.6	.2	—	.2
Total	37				

reliable tests such as mixing with water, tasting, or touching suggests that the student has a limited understanding of the principles that underlie the procedures taught as part of the Mystery Powders unit.

To investigate the strategies students used in carrying out the assessment task, we examined the degree to which students used the most reliable tests (vinegar, iodine, hand lens) for each of the six bags (see Figure 6). Recall that interviewers recorded the sequence of tests students conducted. Students with high scores appeared to invoke the strategy that you gather all possible evidence for each powder before reaching a conclusion. They used vinegar and iodine on all the bags, water on half of the bags, and sensory input for approximately one-third of the bags. This strategy, although systematic, is less efficient than what we might expect from students who understand the principles underlying this set of testing procedures.

In contrast to their higher scoring peers, those with evidence scores from 0-3 tended to conduct one test for each bag. Moreover, these students were more likely than students at all other score levels to rely to a greater extent on unreliable tests. Two of the five students never used vinegar, iodine, or the hand lens for any of the bags. One student relied solely on taste to identify all the powders; the other student relied solely on touch.

As might be expected, students with scores between 4 and 16 showed more variability in their strategies than either the high- or low-scoring students. They tended to conduct one or two tests per bag, some reliable (vinegar, iodine, sight) and some unreliable (water, taste, touch). Students' strategies were generally more effective or efficient at the high end of this score range than at the low end. For example, only two of the nine students with scores between 4 and 7 used more than two tests per bag; students with scores between 8 and 15 used vinegar and iodine as a test for more than one-half of the six bags.

Monitoring

Proficient performance is characterized by monitoring of problem solution and attention to feedback from the task (Glaser, 1991). We considered several types of statements or activities as reflecting monitoring behavior: (a) Refer to Journal—check their hypotheses, observations or conclusions with results of previous investigations recorded in their science journals; (b) Check Options—look at the options or choices given in the instructions thus indicating an effort

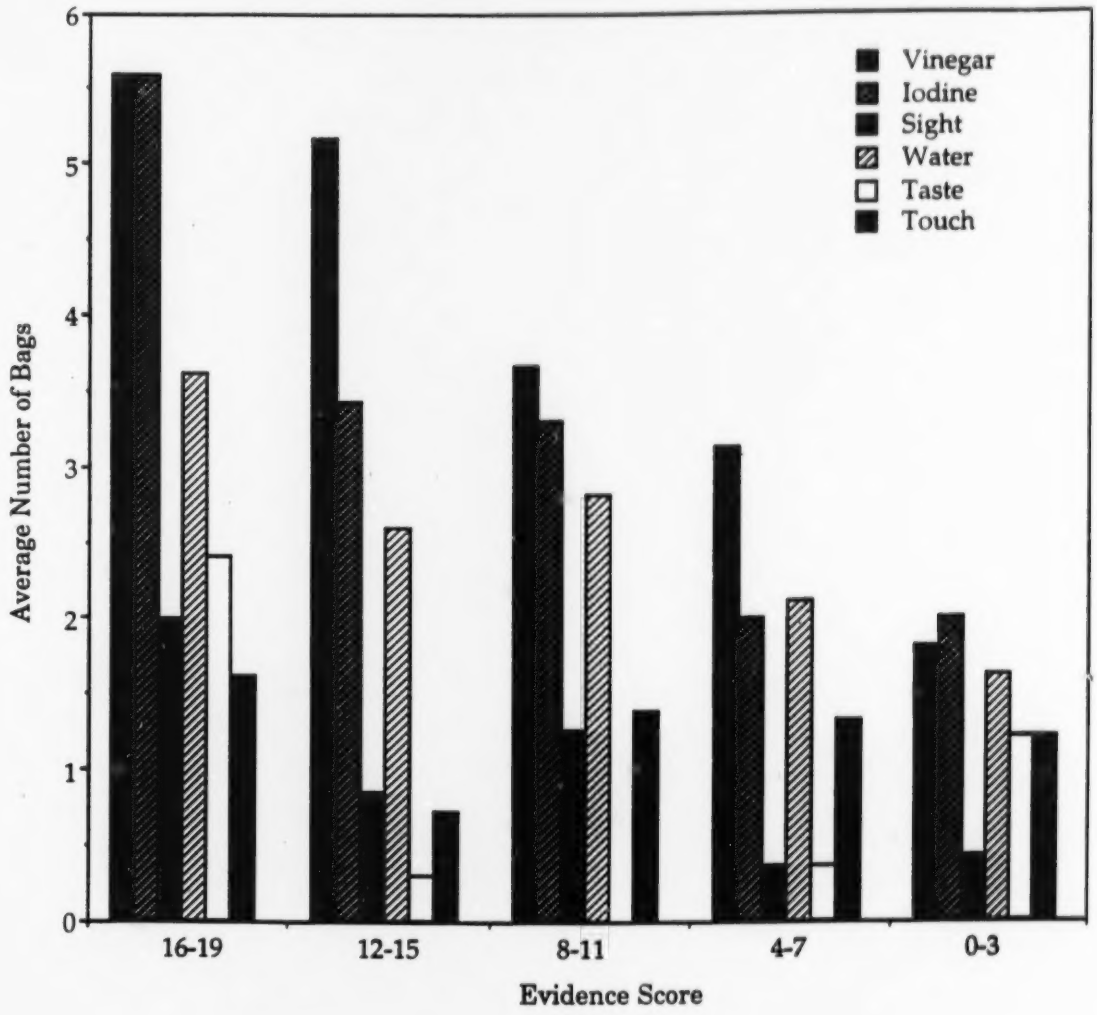

Figure 6. Relationship between evidence score and average use of each test per bag.

to operate within the constraints of the task; (c) Review Answers—check results and conclusions with list of possible options to make sure all were accounted for; (d) Retest Bags—confirm observations or conclusions by retesting a bag; (e) Express Confusion—recognize that a hypothesis was not confirmed or that a test appeared to duplicate findings of another bag thereby suggesting an error might have been made; and (f) Compare Bags—examine tests and observations across bags.

Monitoring by the students with high evidence scores could be differentiated from monitoring by those with low scores in two respects. First, those who scored high made a greater number of self-monitoring statements while carrying out the assessment than those who scored lower. On average, students who scored 12 or more engaged in the six forms of monitoring to a greater extent than those students who scored less than 12 (see Table 4).

Second, students who scored high engaged in more effective forms of self-monitoring than lower scoring students (see Figure 7). Eighty percent or more of the highest scoring students (16-19) consulted their notebooks to compare their tests and observations, checked the list of possible contents for the six bags to check their conclusions, and reviewed their answers to check for matches with the list of options. In addition, approximately 60% of these students compared their tests and observations across the bags, noting similarities and differences, and expressed confusion when their hypothesis did not match the evidence from their investigation. These three forms of monitoring yield immediate, adaptive feedback/information to help students operate within the constraints of the task.

Low-scoring students referred less to their notebooks and seemed to work more from their immediate results. This is reflected in their reliance on comparing across bags (see Figure 7). Moreover, these students relied on their memory of prior activities (e.g., *"I remember we did this in class"*) instead of

Table 4

Average Number of Instances of Various Forms of Monitoring

Evidence score	Refer to journal	Check options	Review answers	Retest bags	Express confusion	Compare bags	Total
20-24	—	—	—	—	—	—	—
16-19	2.8	2.8	0.6	0.4	1.6	1.6	10.0
12-15	3.9	1.7	0.9	1.4	1.6	1.4	10.9
8-11	1.2	0.9	0.2	0.5	0.8	0.6	4.0
4-7	0.0	0.4	0.1	0.4	0.4	0.4	1.9
0-3	1.8	0.2	0.6	0.2	1.2	0.6	4.6

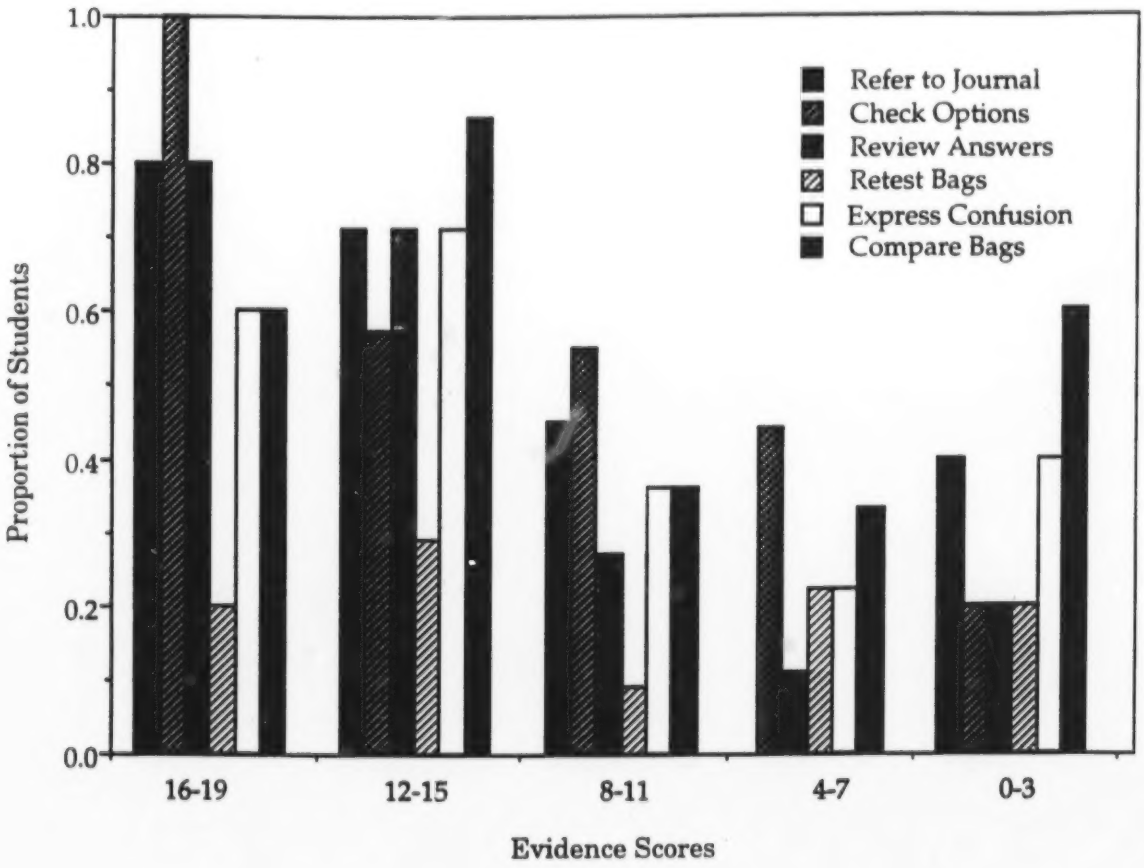

Figure 7. Relationship between score and proportion of students who engage in various forms of monitoring.

looking at a written record of the events that took place in class (referring to their science journals). In fact, students with scores less than 12 relied on this method to a greater extent than on any of the methods of monitoring.

Summary and Conclusions

Assessments developed to support and enhance instruction in hands-on science classrooms ask students to reason with subject matter knowledge to solve a problem. The scoring attempts to focus on the thinking and reasoning process by which the solution is generated, and on key aspects of the

performance drawn from an understanding of the principles underlying the topic. The goal is to provide feedback to students about the extent of their learning and to teachers about the effectiveness of their practice. The degree to which these assessments meet this dual requirement—evaluating student learning and informing instructional practice—is not adequately addressed in traditional concerns for reliability and validity. Evaluation that documents the nature and extent of student thinking and reasoning required for optimal performance is needed.

In this paper we have suggested one possible strategy for examining the correspondence between student score and nature and extent of reasoning and understanding. Student performance was evaluated with respect to four dimensions of proficient performance (planning, content knowledge, strategies, and self-monitoring skills) derived from the cognitive psychology literature on the nature of expertise. It was reasoned that if successful task completion is dependent on higher order thinking skills, then students who score high should generate an initial plan, display generalized content understanding, engage in strategic problem solving, and effectively monitor their performance.

Results indicate that performance and task understanding on the Mystery Powders assessment were very low. More than one-half of the students scored 11 or less out of the 24 possible points. Not surprisingly, then, the most proficient performances were not observed. Nevertheless, high- and low-scoring students displayed qualitatively different performance characteristics. High-scoring students, in general: (a) provided an example of a test and corresponding observation for one powder when asked for an overall plan; (b) demonstrated a generalized understanding of the principles underlying the unit; (c) displayed a systematic approach to solving the problem by gathering all possible information before drawing conclusions; and (d) engaged in effective and flexible monitoring of their performance by referring to their prior investigations (e.g., looked at their journal), and operating within the constraints of the task (e.g., checked list of options).

In contrast, the low-scoring students' plans consisted of restating the problem or naming the equipment they would use. These students believed they could identify all substances with just one test. Their trial-and-error strategy was to do a test and see what happens. In monitoring their

performance, they relied primarily on their memory of prior classroom activities or comparison of current observations regardless of their relevance.

Describing these characteristics of students with respect to their performance scores highlights the details of proficiency acquired with this unit. For example, as demonstrated with the analysis presented here, students for the most part have learned a set of procedures for investigating the properties of powders. However, they have little understanding of the principles underlying those procedures. Even the most proficient students demonstrated a systematic but not efficient approach to solving the problem. Second, the majority of the students did not recognize the value of their science journals as an effective and efficient strategy for monitoring their performance. Rather, they relied on their memory (often incorrectly) of their classroom experiences as a check on their hypotheses and conclusions in the assessment situation.

Efficient, principled problem solving and effective, flexible monitoring are the hallmarks of competent performance. Teachers can support students in developing these characteristics by drawing attention to the overall goals of the unit as they attend to the procedures they are carrying out. Second, they can encourage students to reflect on the effectiveness of others' strategies, and on how what they are doing relates to what they are trying to accomplish. Third, teachers can design classroom activities to make critical cognitive activity as relevant and visible as possible to enable students to see the effectiveness of these activities. Awareness of and attention to these sorts of activities, which differentiate more from less proficient performance, can support the development of thinking and reasoning in the elementary science classroom.

References

Baker, E. M., O'Neil, H. F. Jr., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist, 48*(12), 1210-1218.

Baxter, G. P., Glaser, R., & Raghavan, K. (1993). *Analysis of cognitive demand in selected alternative science assessments.* Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Chi, M. T. H., Glaser, R., & Farr, M. (Eds.). (1988). *The nature of expertise.* Hillsdale, NJ: Erlbaum.

Ericsson, A. K., & Simon, H. A. (1984). *Protocol analysis. Verbal reports as data.* Cambridge, MA: MIT Press.

Glaser, R. (1991). Expertise and assessment. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 17-30). Englewood Cliffs, NJ: Prentice Hall.

Glaser, R. (1992). Expert knowledge and processes of thinking. In D. Halpern (Ed.), *Enhancing thinking skills in the sciences and mathematics* (pp. 63-75). Hillsdale, NJ: Erlbaum.

Schauble, L., Glaser, R., Duschl, R., Schulze, S., & John, J. (in press). Students' understanding of the objectives and procedures of experimentation in the science classroom. *The Journal of the Learning Sciences.*